90 La variable Khi- carré

90.1 Distribution de Khi -Carré

http://www.fundp.ac.be/biostats/biostat/modules/module90/index.html - module_90

90.1.1 Utilité

La variable Khi -Carré est un modèle qui exprime la distribution de sommes de carrés d'écarts standardisés.
Elle permet de calculer la probabilité d'observer des écarts dus au hasard entre des fréquences observées et celles prévues par une loi de probabilité.
Elle permet également de décrire la distribution de la variance des échantillons pris dans une population.
C'est avec la distribution normale un des outils les plus utilisés en statistiques biomédicales.

90.1.2 Principe

Imaginons un modèle qui répartisse les observations en deux catégories, par exemple mâles et femelles, dans une population de sex -ratio[1] 0,5.

P(mâle) = 1/3, P(femelle) = 2/3; ratio = 0,5

Comptons la fréquence des mâles et des femelles dans un échantillon (n = 87) et la fréquence théorique attendue suivant la répartition 1/3, 2/3.

Calculons un écart quadratique entre les fréquences observées et théoriques, standardisé par la fréquence théorique :

$$\frac{(f_{obs} - f_{th})^2}{f_{th}} : \frac{(23-29)^2}{29} = 1.24$$

et rassemblons les valeurs dans un tableau :

	mâles	femelles	total
f_i observée	23	64	87
f_i théorique	29	58	87
écart quadratique standardisé	1.24	0.62	1.86

Tableau 90 -90-1 Fréquences observées et théoriques, écart quadratique, standardisé par la fréquence théorique, totaux.

[1] *Sex -ratio : rapport numérique des sexes à la naissance (mâles/femelles). (Dans l'espèce humaine il est d'environ 105 garçons pour 100 filles soit 1.05).*

Les fréquences observées fi correspondent approximativement à la valeur de X = Po(μ), expression dans laquelle $\mu = np_i$ avec n la taille de l'échantillon et p_i la probabilité d'appartenir à la catégorie i.

La variance de cette fréquence observée est donc Var(X) = $\mu = np_i$.

La quantité $\frac{(f_{obs} - f_{th})}{\sqrt{f_{th}}}$ est donc approximativement une variable Z(0 ;1).

L'écart global entre les observations et le modèle est calculé par la statistique[2] :

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(f_{i\ obs} - f_{i\ th})^2}{f_{i\ th}}$$ **Équation 90-1**

qui suit approximativement une distribution théorique

$$\chi^2_{k-1} = \sum_{i=1}^{k} Z_i^2$$

expression dans laquelle k représente le nombre de catégories et k -1 le nombre de degrés de liberté, dont dépend la forme de la courbe.

Si l'on répète l'observation un grand nombre de fois, on obtiendra différentes fréquences, et différentes valeurs de χ^2_{obs} .

Echantillon N°	mâles	femelles	χ^2_{obs}
1	23	64	1,860
2	29	62	0,088
3	25	60	0,588
4	25	73	2,699
5	32	63	0,005
6	37	66	0,311
7	32	74	0,472

Tableau 90 -90-2 Répétition de l'expérience, fréquences observées et valeurs de χ^2_{obs}.

Comparons les valeurs obtenues pour χ^2_{obs} et χ^2_{k-1}

[2] χ *: vingt -deuxième lettre de l'alphabet grec, se prononçant khi. Dans les références statistiques, s 'écrit aussi chi -carré,khi -deux….*

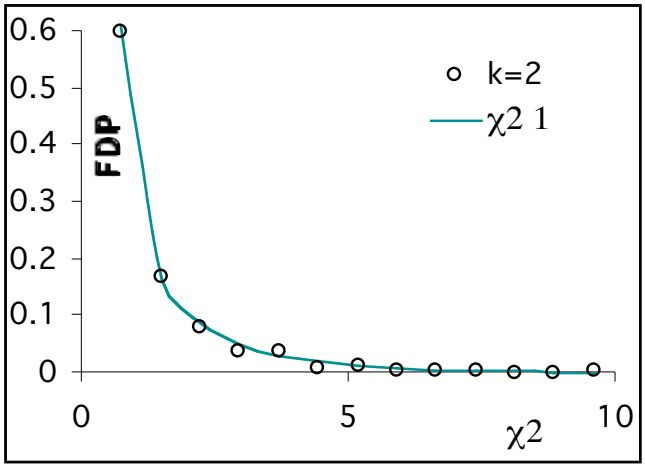

Figure 90 -1 Comparaison des écarts quadratiques standardisés à la distribution théorique de khi -carré avec un degré de liberté.

90.1.3 Exemple

Imaginons un modèle qui répartisse les observations en trois catégories, par exemple les produits AA, Aa (ou aA) et aa de plusieurs croisements hétérozygotes, de probabilité 25%, 50% et 25% respectivement. Effectuons 5 fois l'expérience qui consiste à relever la fréquence de chaque génotype[3].

N°	AA	Aa	aa	$\chi 2$obs
1	27	49	20	1,06
2	17	53	31	4,13
3	27	46	22	0,62
4	22	46	27	0,62
5	28	53	17	3,12

Tableau 90 -90-3 Répétition de l'expérience, fréquences observées et valeurs de χ^2_{obs}.

[3] *Génotype : patrimoine génétique d'un individu dépendant des gènes hérités de ses parents. Phénotype : (du grec. phainein, montrer, et tupos, marque) : ensemble des caractères somatiques apparents d'un individu, qui expriment l'interaction du génotype et du milieu.*

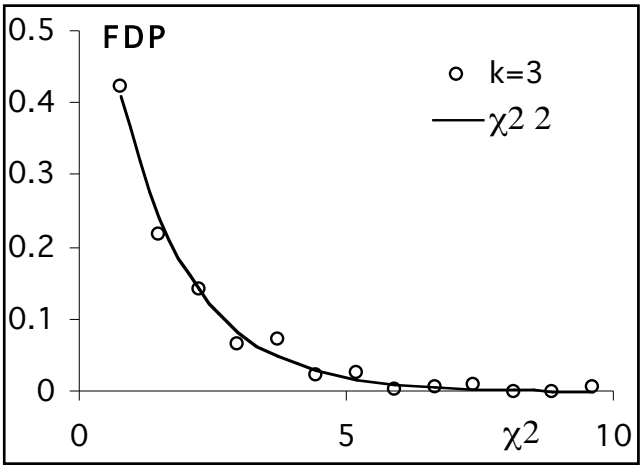

Figure 90 -2 Comparaison des écarts quadratiques standardisés à la distribution théorique de khi -carré avec deux degrés de liberté.

Considérons la probabilité a priori de 10 acides aminés[4] de se trouver dans une hélice alpha[5] et dénombrons leur fréquence dans 4 protéines.

Acide aminé	**Proba - bilite**	**Protéines**			
		1	**2**	**3**	**4**
Ile	0,03	4	3	3	3
Asn	0,05	8	5	8	4
Val	0,07	12	7	8	4
Thr	0,07	5	8	7	9
Tyr	0,07	9	7	10	5
Leu	0,13	8	15	15	12
Pro	0,13	14	13	12	13
Glu	0,15	9	18	16	17
Gly	0,15	17	12	16	10
Met	0,15	12	16	18	20
total	**1**	**98**	**104**	**113**	**97**
χ^2_{obs}		12,32	1,49	2,35	6,40

Tableau 90 -90-4 Probabilités, fréquences observées et χ2 observé pour 10 acides aminés répertoriés dans les hélices alpha de 4 protéines.

[4] *Acide aminé : substance organique ayant une fonction acide et une fonction amine. Vingt acides aminés sont les constituants fondamentaux des protéines.*

[5] *Hélice alpha : structure secondaire d'une protéine, plus ou moins longue, dans laquelle les acides aminés forment un angle caractéristique d'un pas d'hélice.*

Si l'on répète l'expérience sur un plus grand nombre de protéines, on observe une distribution de χ^2_{obs} qui peut se comparer à une distribution théorique de χ^2 avec 9 degrés de liberté.

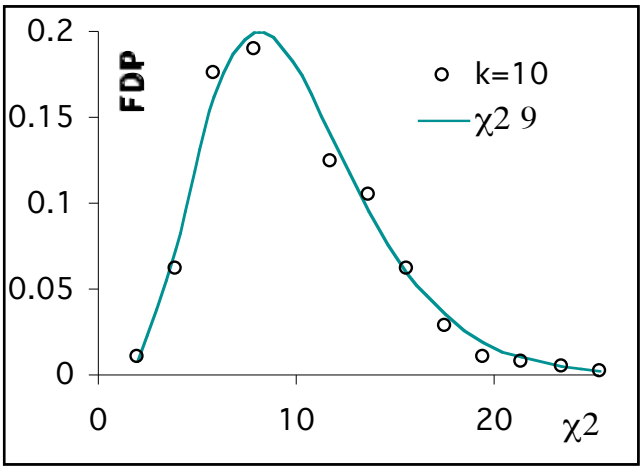

Figure 90 -3 Comparaison des écarts quadratiques standardisés à la distribution théorique de khi -carré avec 9 degrés de liberté.

90.1.4 Tables et graphiques

La variable χ^2 est utilisée pour décrire la distribution de sommes des carrés des écarts.

Les degrés de liberté déterminent la forme de la courbe et dépendent du nombre de catégories dans lesquelles les fréquences sont dénombrées.

Plus le nombre de degrés de liberté augmente, plus χ^2 tend vers une v.a. Normale.

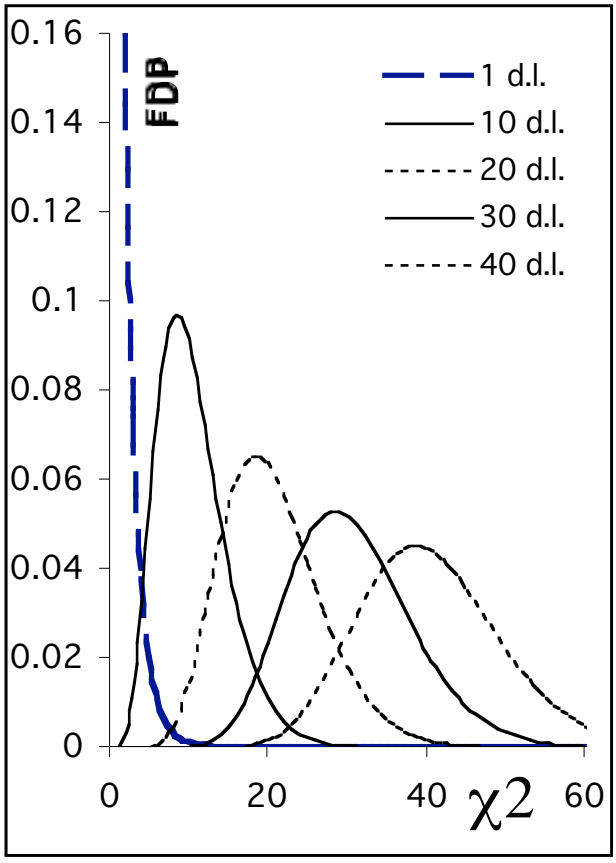

Figure 90 -4 Comparaison de fonctions χ^2 avec différents nombres de degrés de liberté. La distribution de χ^2 avec un petit nombre de degrés de liberté est fortement asymétrique.

La table de χ^2 est généralement présentée de la façon suivante :

	0,9	0,95	0,975	0,99
1	2,71	3,84	5,02	6,63
2	4,61	5,99	7,38	9,21
3	6,25	7,81	9,35	11,34
4	7,78	9,50	11,14	13,28

Tableau 90 -90-5 Extrait de la table de χ^2
En tête de colonne, les probabilités π , en tête de ligne, les degrés de liberté (k). Chaque case comprend $P(\chi^2_\kappa \leq \chi^2_{\kappa;\pi})$.

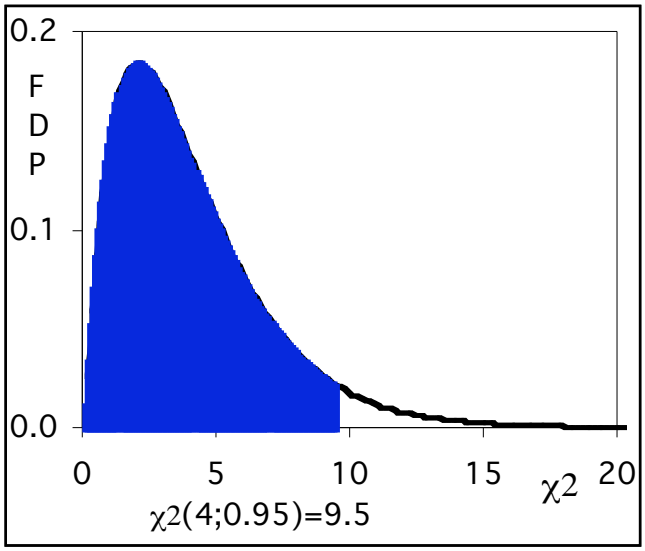

Figure 90 -5 Illustration de la probabilité reprise dans la table, 4 d.l. π= 0.95.

90.1.5 Dans le tableur Excel

la fonction LOI.KHIDEUX(x;dl) renvoie

$P(\chi^2 dl \geq x)$.

Dans notre exemple,
1 -LOI.KHIDEUX(9.5 ;4) renvoie 0,95.

90.2 Distribution d'échantillonnage de S^2

90.2.1 Utilité

La variabilité des mesures est le point de référence qui permet de juger de la différence entre les moyennes de différents échantillons.
La distribution d'échantillonnage de la variance, basée sur la distribution de Khi -carré, permet de calculer la probabilité d'observer une variance donnée dans une population.

Il est essentiel de pouvoir comparer des variances pour juger de ce qu'un traitement, par exemple, a modifié la moyenne d'une population ou sa variance.

> *Imaginez qu'un éleveur de poulets consente une dépense supplémentaire pour un aliment enrichi d'un stimulateur de croissance. Si la substance modifie la variance du poids au lieu de modifier sa moyenne, il n'aura rien gagné globalement. Au contraire, la variabilité du poids sera un obstacle au conditionnement pour la consommation.*

90.2.2 Redéfinition de la variance

Lorsque la variance de l'échantillon doit représenter celle de la population, sa définition est un peu différente de celle de l'écart quadratique moyen vu en statistiques descriptives. En effet, S_n^2 = SCE/n représente la variance de l'échantillon, mais σ^2 est $E(S_{n-1}^2)$ exprimé de la façon suivante :

$$S_{n-1}^2 = \frac{SCE}{n-1}$$ **Équation 90-2**

> *Intuitivement, on peut comprendre la nuance de la façon suivante : supposons un échantillon d'une seule observation : S_n^2 est nulle. On ne peut évidemment pas en déduire que la variance de la population est nulle : elle est en fait indéterminée, la notion de dispersion n'ayant pas de signification vis -à -vis d'une valeur unique. Par contre, suivant cette nouvelle définition, si l'échantillon ne contient qu'une valeur, S_{n-1}^2 = 0/0, ce qui est indéterminé.*

Cette distinction n'est importante que lorsque l'échantillon est petit, dans le cas contraire il n'y a pas grande différence entre n et n -1.

Dans le contexte des bio -statistiques, nous avons très souvent de petits échantillons et presque toujours en référence à la variance de la population. Par défaut, il vaut donc mieux utiliser systématiquement la définition S_{n-1}^2.

Les deux définitions S_n^2 et S_{n-1}^2 cohabitent sans qu'une convention de notation bien établie les distingue : on notera généralement S_x^2 pour désigner que l'on parle de la variance de la variable X ou plus simplement S^2 par facilité dans l'utilisation des symboles.

Dans le tableur Excel, S_n^2 se calcule par var.p() et S_{n-1}^2 par var(). Sur les calculettes, les symboles $S_n^2, \sigma_n^2, \sigma_{n-1}^2$, VAR, S.D. … cohabitent, au point que parfois, la variance n'est pas le carré de l'écart -type, l'une étant calculée en référence à n et l'autre en référence à n -1 !

90.2.3 Distribution de la variance

Pour visualiser la distribution d'échantillonnage, envisageons la mesure de la variance de 5 observations prises au hasard dans une population de variance σ^2, réalisée 1000 fois. La distribution de S^2 suit une distribution asymétrique :

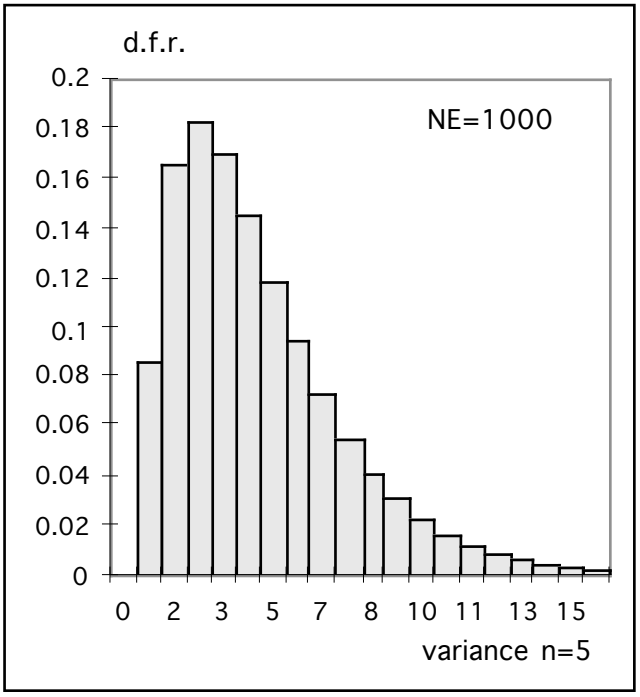

Figure 90 -6 Distribution de densité de fréquences relatives(d.f.r) de S^2 calculées sur 1000 échantillons (NE) de 5 observations.

En considérant la SCE = (n -1)S^2 , le rapport

$$\frac{SCE}{\sigma^2} = \left(\frac{x_1 - Mx}{\sigma}\right)^2 + \left(\frac{x_2 - Mx}{\sigma}\right)^2 + \ldots$$

correspond à la somme des carrés d'écarts standardisés, soit une somme de Z^2, ce qui correspond à la définition théorique d'une variable Khi -carré avec n - 1 degrés de liberté.

La distribution du rapport $\frac{(n-1)S^2}{\sigma^2}$ suit une distribution χ^2_{n-1}

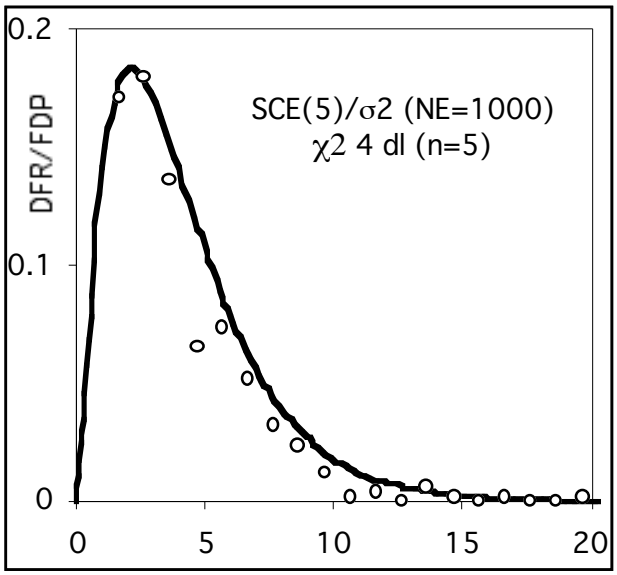

Figure 90 -7 Distribution de densité de fréquences relatives (DFR) observée du rapport SCE/σ2 calculé sur 1000 échantillons de 5 valeurs, superposée à la fonction de densité de probabilité (FDP) d'un v.a. χ2 avec 4 degrés de liberté.

90.2.4 Exemple

On peut donc utiliser la distribution de χ^2 pour établir la probabilité d'observer une variance S^2 donnée dans un échantillon de taille n prélevé dans une population de variance σ^2 connue.

Supposons que la variabilité du taux de progestérone soit bien établie dans une population de brebis : $\sigma^2 = 40$.

Quelle est la probabilité d'obtenir dans un échantillon de 3 individus une variance supérieure à 100?

$$\frac{(n-1)S^2}{\sigma^2} = \frac{2 \times 100}{40} = 5$$

Les tables ne fournissent pas de valeur précise pour $\chi^2_2 > 5$. On y trouve $P(\chi^2_2<6)=0,95$ et $P(\chi^2_2<4,5)=0,9$. $P(\chi^2_2>5)$ est donc comprise entre 0,10 et 0,05.